An Infrastructure for Passive Network Monitoring of Application Data Streams

Brian L. Tierney, Deb Agarwal,
Goujun Jin, José Maria González

Lawrence Berkeley National Laboratory

U.S. Department of Energy
Office of Science

Self Configuring Network Monitor (SCNM)

BERKELEY LAB

- **SCNM is a passive monitoring system designed to address the following issues:**
  - **Ability for network *users* to monitor their traffic as it traverses the network**
    - SNMP provides a very coarse level of monitoring, and end users rarely have permission to access this data
    - DAG-based systems provide packet capture ability, but end users don't have access to this data
  - **Ability to identify the source of network congestion or other problems (e.g: is packet loss a LAN or WAN issue)**
  - **Ability for application developers to characterize their own traffic, and how it is impacted by the network (e.g.: Jitter)**
  - **Protocol analysis and debugging**
    - Often not possible to capture packet traces at the sending host
      - tcpdump will often lose packets when trying to capture a high bandwidth stream

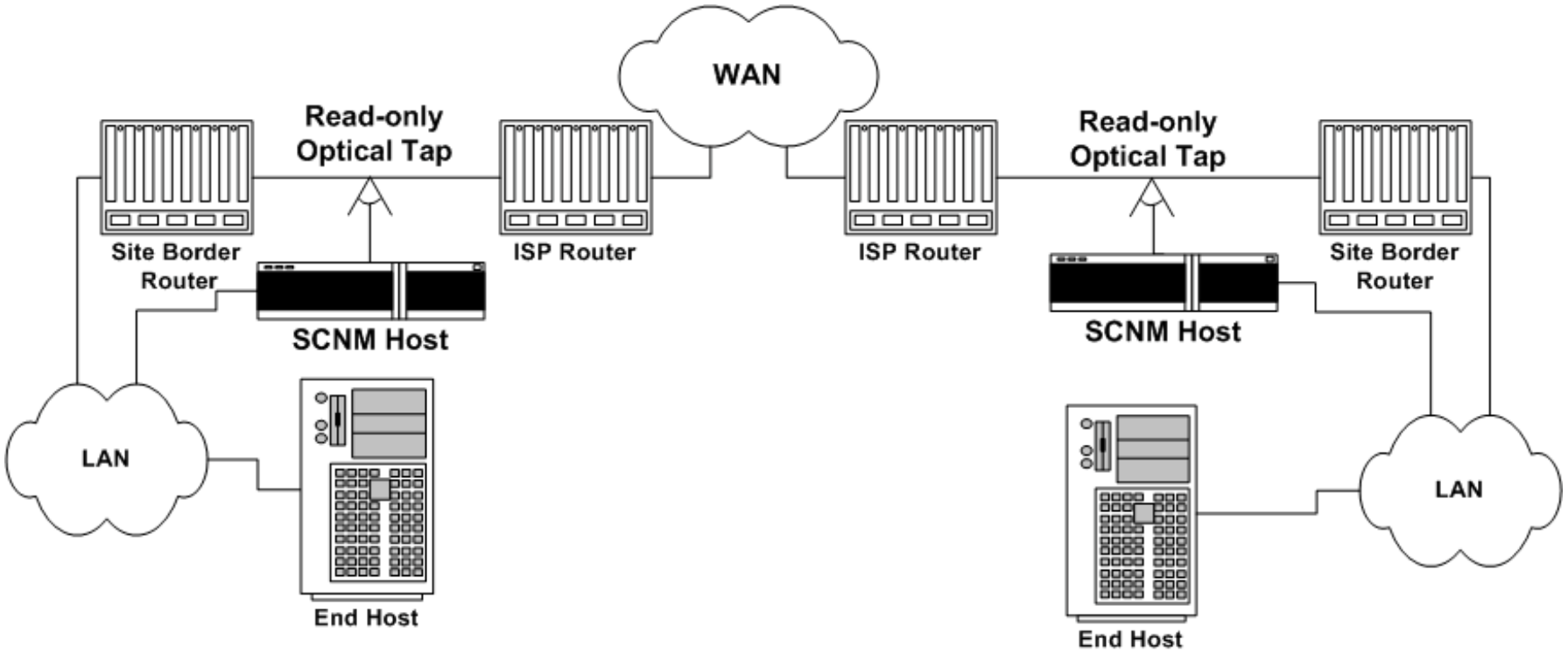
U.S. Department of Energy
Office of Science

Self Configuring Network Monitor (SCNM) Goals

- Based on low cost, “off the shelf” components
- Easy to activate
- Secure
  - Users can monitor their own data, without the need to involve network administrators
- Easy to deploy, configure, and administer

U.S. Department of Energy
Office of Science

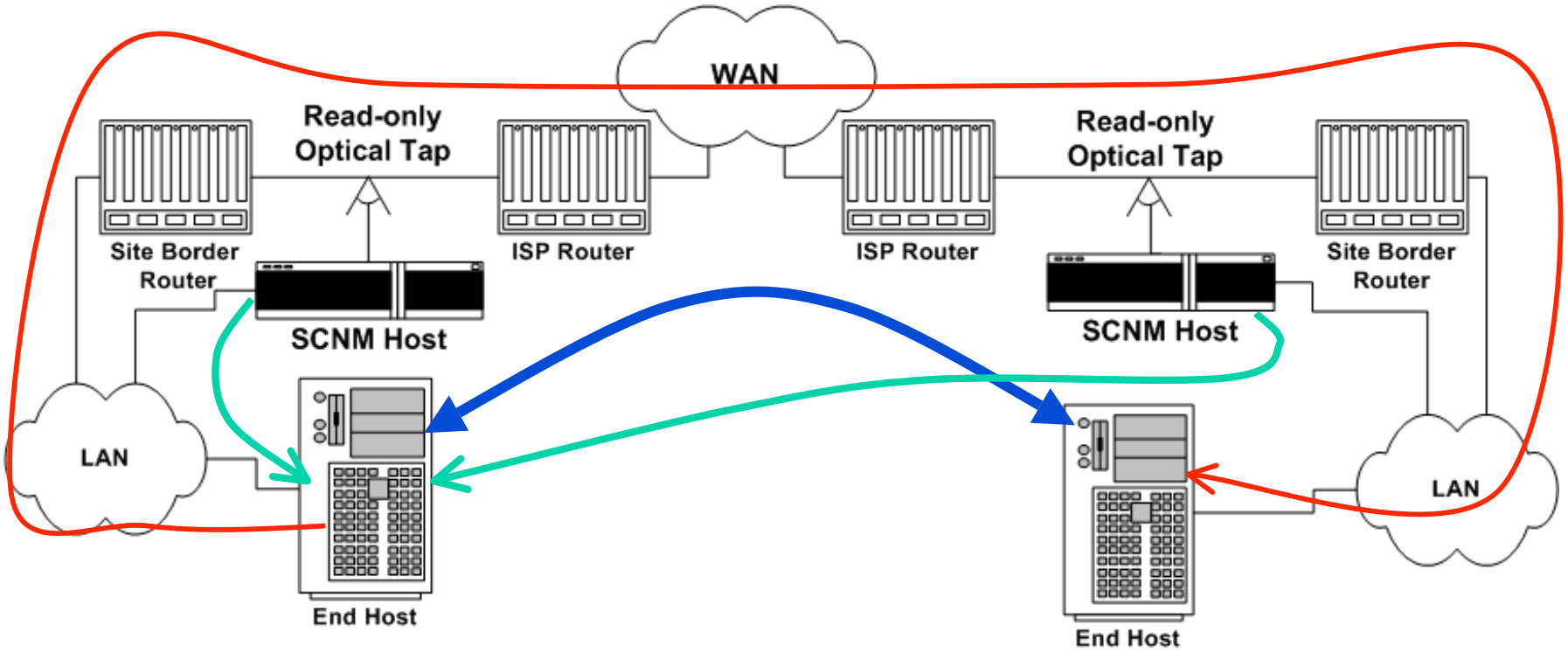
Typical Deployment



U.S. Department of Energy
Office of Science

Monitor Activation

BERKELEY LAB



Data Flow

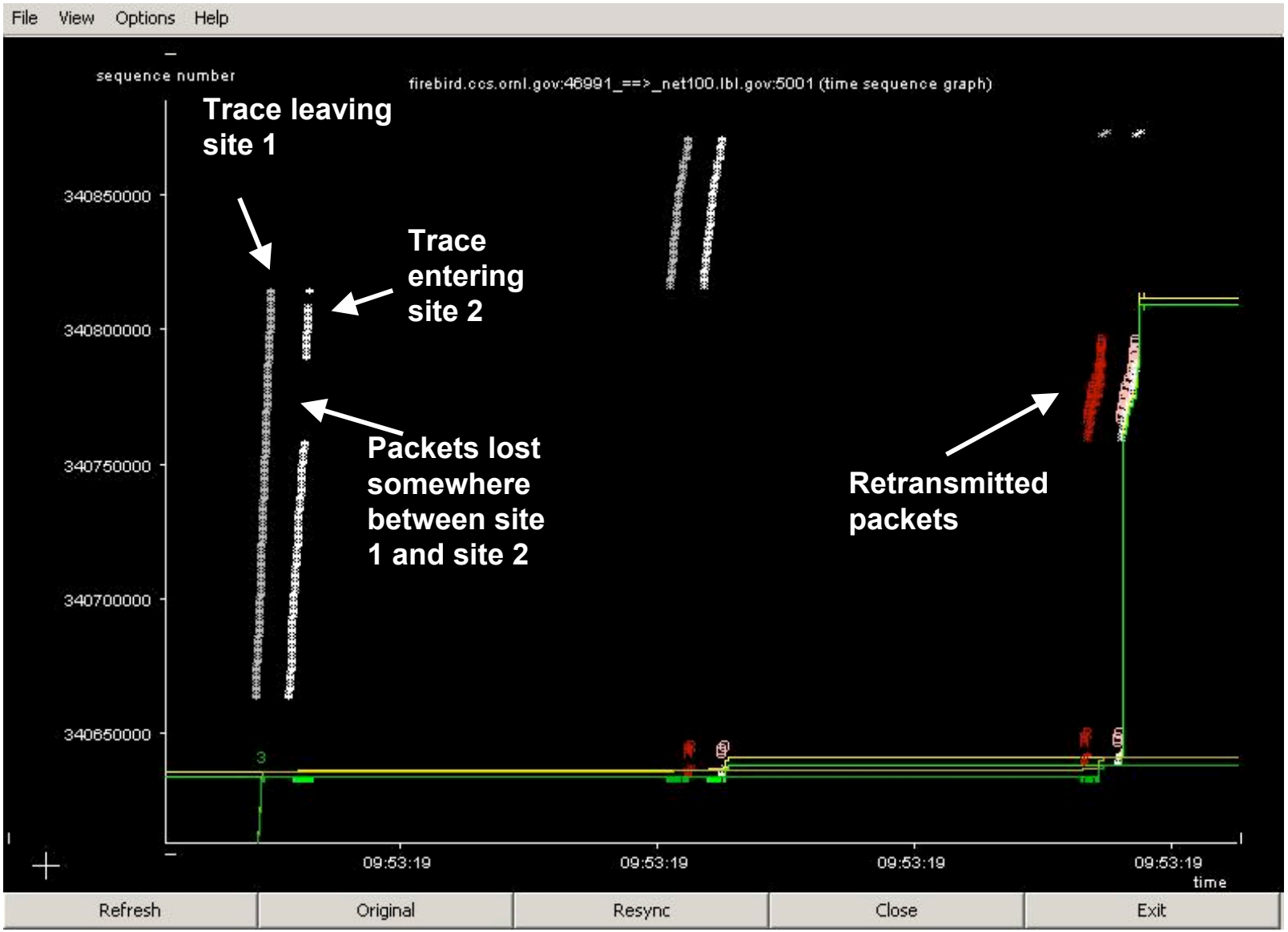
Activation packet

Packet Headers

U.S. Department of Energy
Office of Science

Typical Result

BERKELEY LAB



U.S. Department of Energy
Office of Science

Passive Monitor Host

BERKELEY LAB

- Installed at critical points in the network (i.e.: next to key routers)
- Uses read-only 1000BT fiber tap (e.g.: from Netoptics)
- Passively captures packet headers of monitored traffic
  - Daemon based on libpcap (used in NIMI)
- Configured and activated by application end-points
  - Without network administrator involvement
- Has an additional network interface (usually on an internal network) used for administering the SCNM host and sending monitoring output data

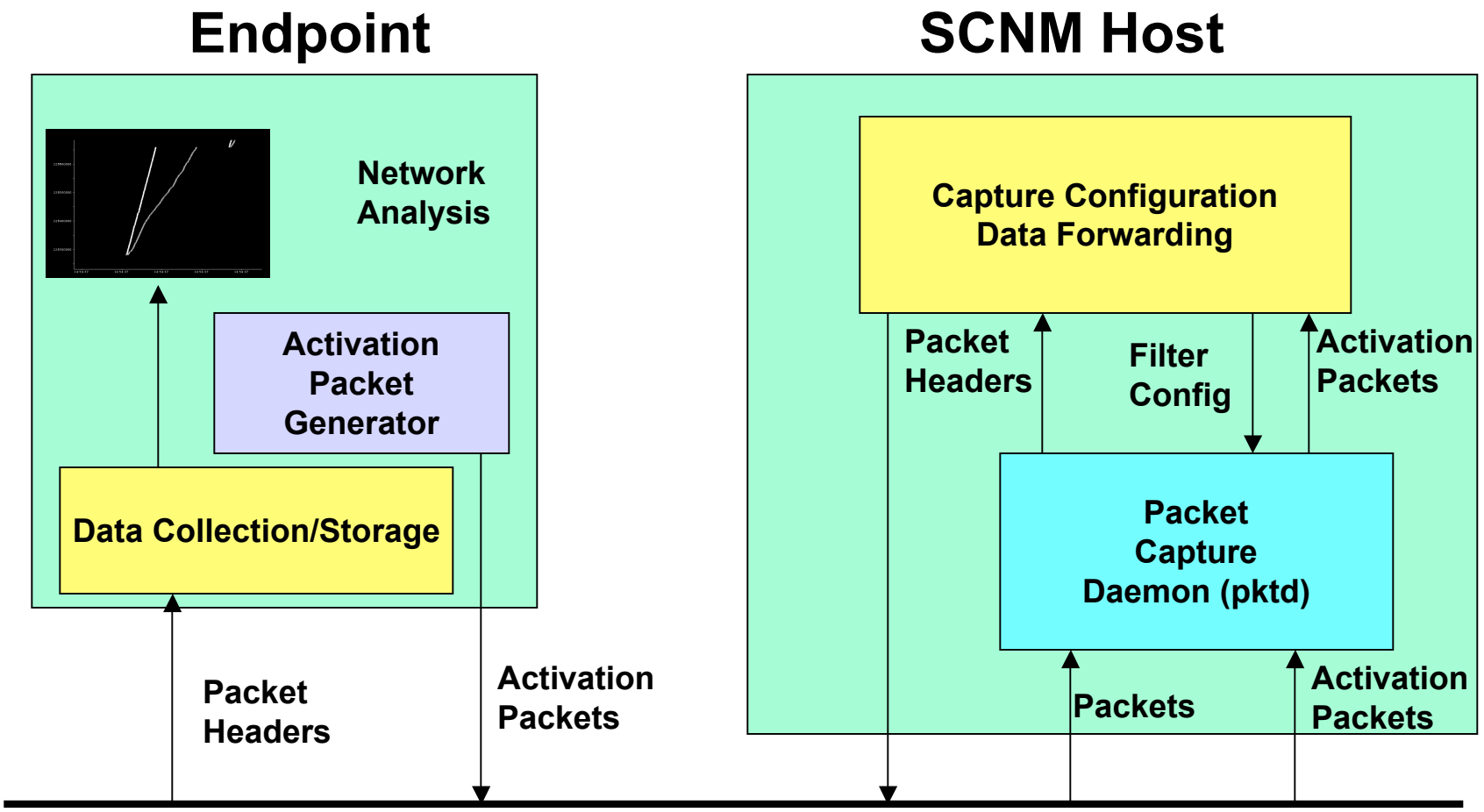
Activation and Configuration

- Activation packets are sent from the application endpoint using UDP and a well known port
  - Sent to the other endpoint of application traffic
  - All monitors along the data path capture activation packet
- Activation packets specify which traffic to monitor
  - Must be the traffic for the same source/dest
- The monitor configures itself to monitor the traffic
- Activation packets are resent periodically to refresh monitor state
- Monitor times out if no activation packets are received

U.S. Department of Energy
Office of Science

System Design

BERKELEY LAB



U.S. Department of Energy
Office of Science

Security Model

BERKELEY LAB

- Monitor host system installed and maintained by network administrators
- A user is only allowed to monitor their own data
  - Or other streams from the same sending host
- The activation packet must be traveling between the source and destination of the traffic to activate monitoring
- The SCNM host will only send resulting packet headers back to the source of the activation packet
- SCNM host logs all monitoring requests
- Network Admin mode (future):
  - Use signed and authorized activation packets to allow the ability to:
    - Activate monitoring from a host that is not one of the endpoints
    - Send packet headers to a host that is not one of the endpoints

U.S. Department of Energy
Office of Science

Implementation Issues

BERKELEY LAB

- Currently deployed packet capture hosts (almost 2 years old HW)
  - FreeBSD
  - SuperMicro 370DL motherboard
  - single P-III Xeon 933MHz CPU
  - PCI 64bit/66MHz IO bus, 133 MHz DIMMs
  - 2 SysKonnect PCI 64 bit 1000BT NICs
    - one for capturing packets in each direction.
- Many performance issues had to be addressed, including:
  - Interrupt moderation
  - Timestamping packets in the NIC
  - Reducing kernel memory copies
- Using this hardware we have monitored TCP flows of up to 940 Mbits/sec for several minutes

U.S. Department of Energy

Office of Science

Interrupt Moderation Issues

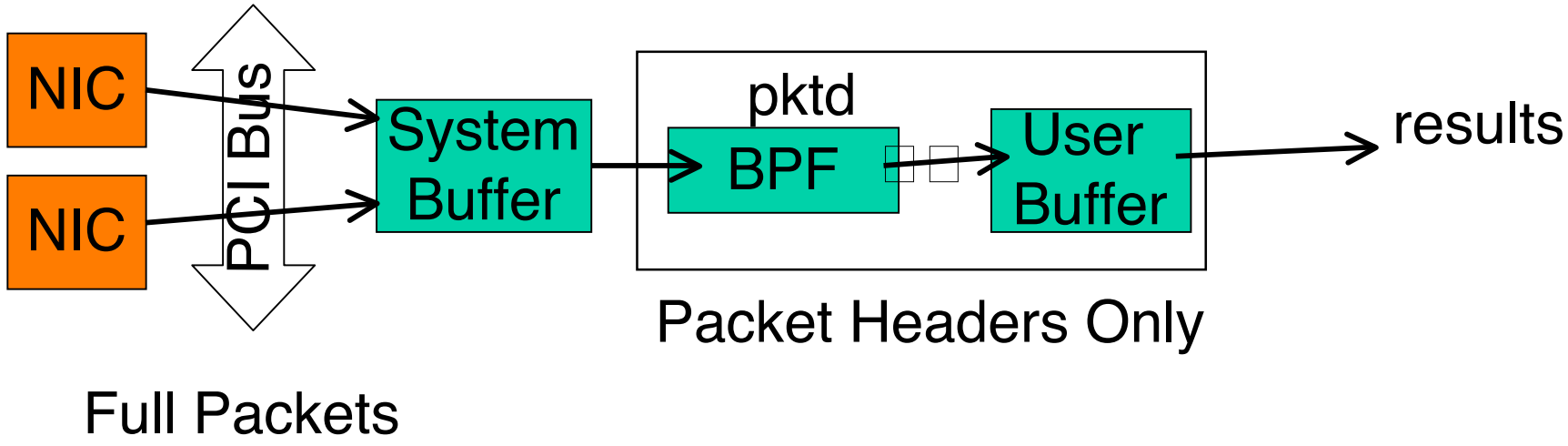
BERKELEY LAB

- Interrupt moderation helps reduce the CPU requirements, but:
- Interrupt moderation period must be short enough to:
  - Avoid running out of kernel receive buffer descriptors
  - Avoid large delays in packet processing
- Drawback of interrupt coalescing:
  - Kernel is no longer able to assign accurate timestamps to the arriving packets.
- Some network cards (including SysKonnect) have an onboard timestamp register
  - Can provide information on the exact packet arrival time, and pass this timestamp to the system buffer descriptor
- We have modified the FreeBSD SysKonnect driver and BPF implementations to use the NIC timestamp, instead of the system clock timestamp
  - We are currently using a 1 ms interrupt moderation period (default is 200 µs)

U.S. Department of Energy
Office of Science

Memory Bandwidth Issues

BERKELEY LAB



- **Memory bandwidth is the most critical HW factor in data capture**
- **Assuming bidirectional traffic, each of the 2 NICs transfer data to the system bus at up to 1 Gbit/s = 125 MBytes/second**
- **Memory bandwidth required is:**
  - **2 x 125 MBytes/sec + around 10% for multiplexing and forwarding packet headers ≈ 300 MBytes/sec**

U.S. Department of Energy
Office of Science

Memory Issues: Current PC Hardware

- Memory bandwidth is up to around 1 GByte/second
  - (e.g: ServerWorks P4DL6 motherboard; 466 MHz DDR memory; 64-bit/133MHz PCI-X bus, 10 GigE NIC)
  - Bottleneck is both the PCI-X bus and the DDR memory
- This system should be capable of capturing almost 4 Gbits/sec (500 MBytes/sec) of bidirectional traffic:
  - 2 x 500 MBytes/sec + 10% needs 1100 MBytes/sec of memory bandwidth
- This is a hardware requirement
  - There are also a number of software tuning issues that must be addressed to achieve this capability
- Will begin testing this configuration soon

U.S. Department of Energy
Office of Science

Header Compression in *pktd*

BERKELEY LAB

- Even though we are only capturing headers, the amount of data collected is quite large
  - For example, a 10 second capture of a 300 Mbps TCP stream contains over 20 MB of header data.
- To reduce this, we have implemented packet header compression
  - Based on the approach used by CSLIP
- Works quite well
  - Achieves an average lossless compression ratio of about
    - 6:1 for UDP headers
    - 4:1 for TCP headers.

Visualization / Analysis Tool

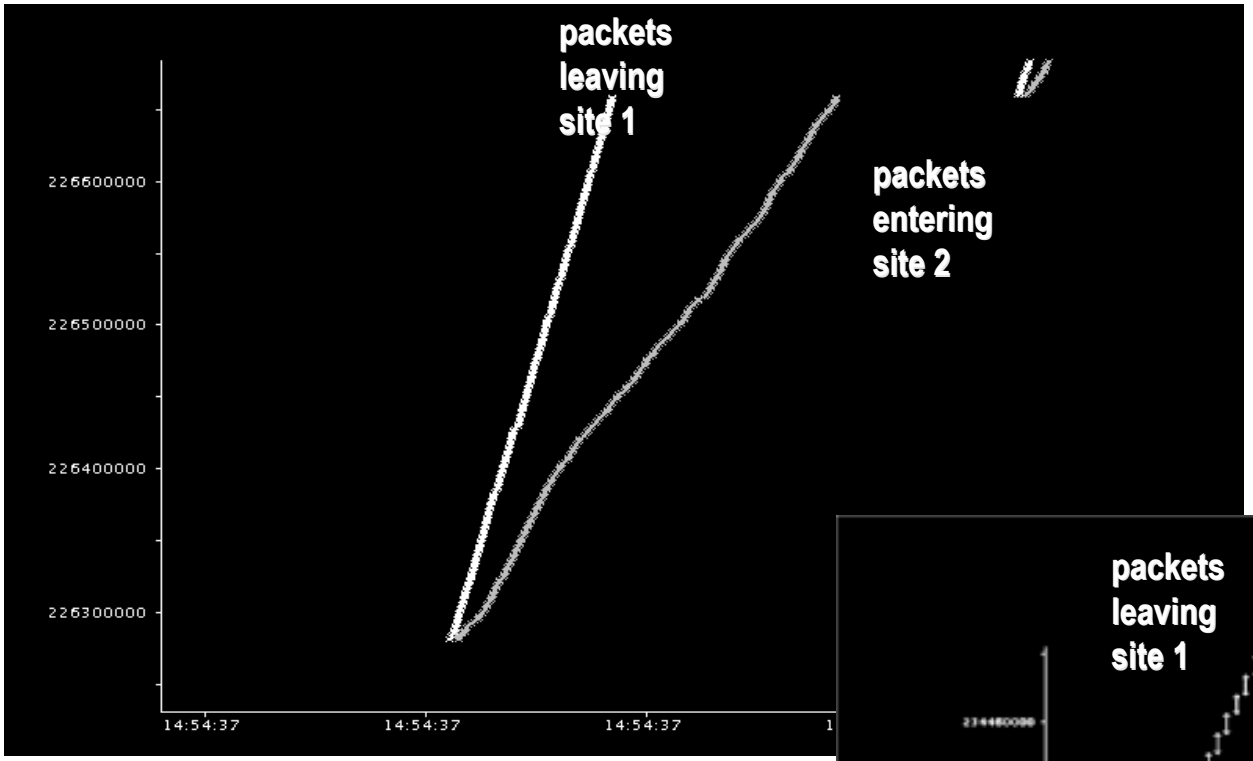
- **SCNMPlot**
  - **Uses tcptrace files for input**
    - http://www.tcptrace.org/
  - **Enhanced version of xplot tool**
    - Allowed multiple input files
    - “Nudge” feature

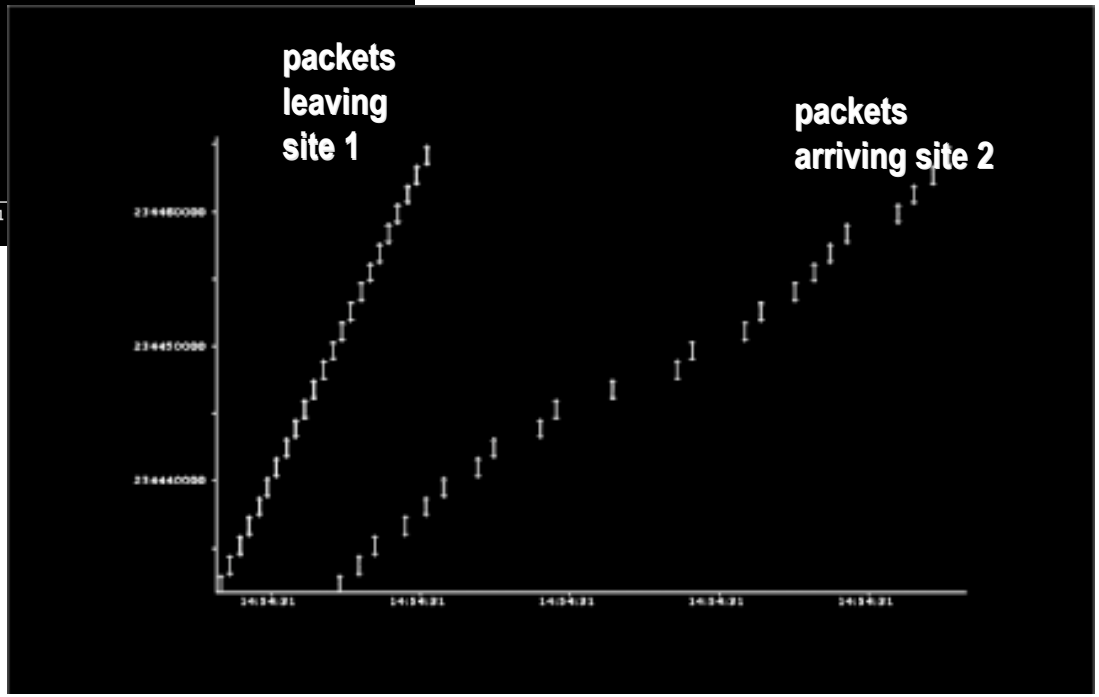
U.S. Department of Energy

Office of Science

Observing Packet Jitter

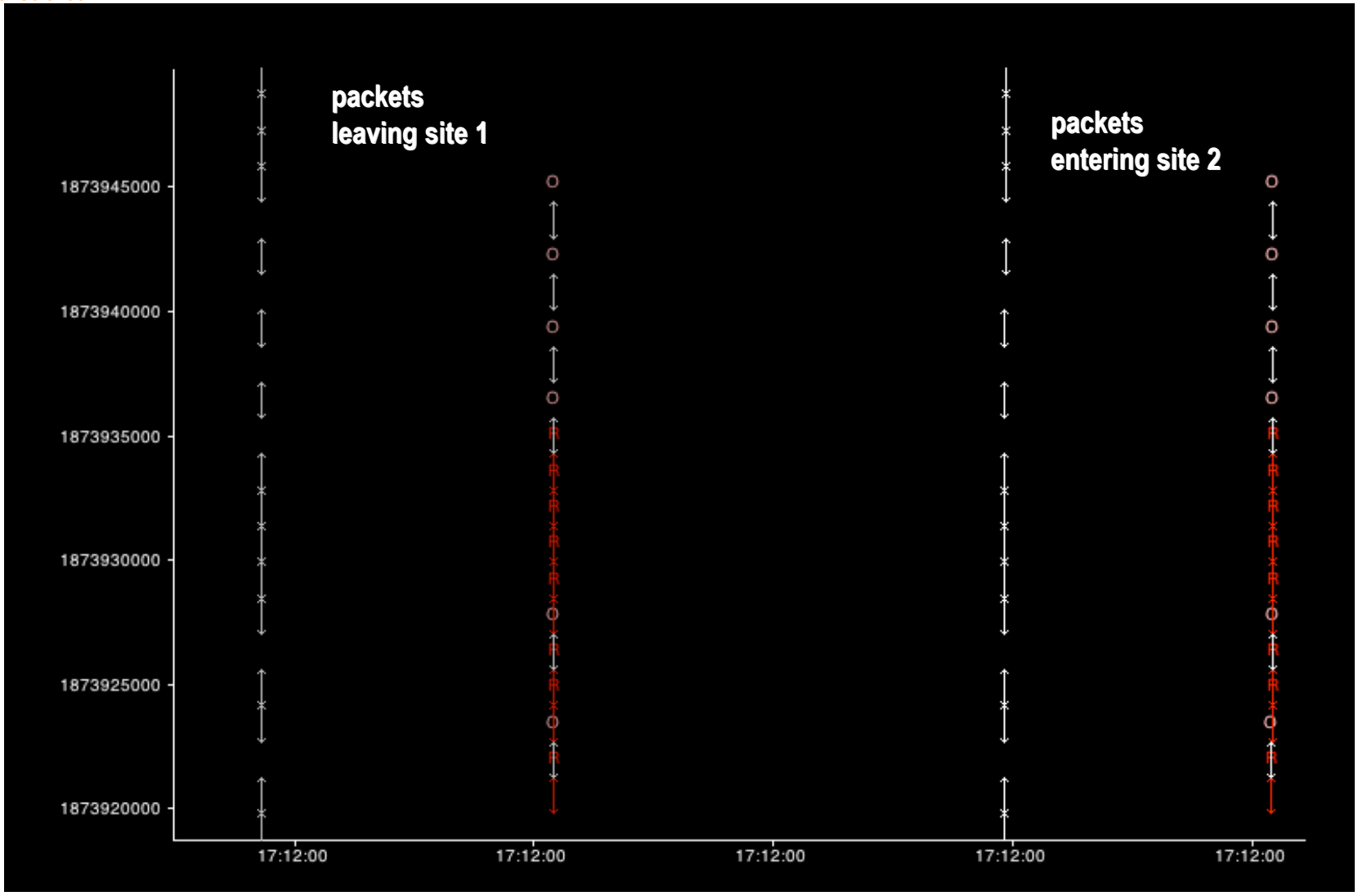
BERKELEY LAB





U.S. Department of Energy
Office of Science

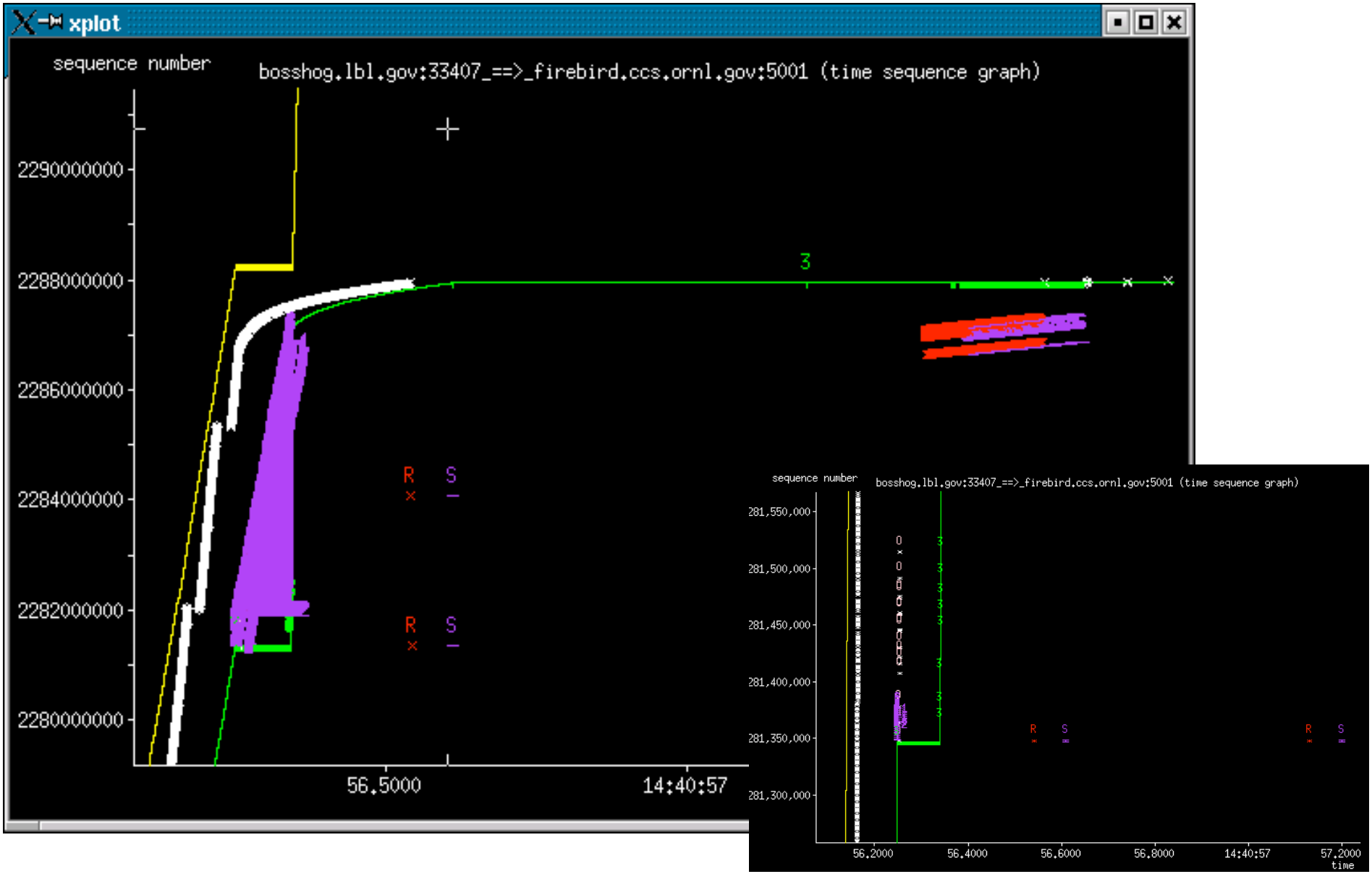
Results: Loss on Sender and Receiver LAN



U.S. Department of Energy
Office of Science

Results: Linux SACK Bug (not seen in local tcpdump)

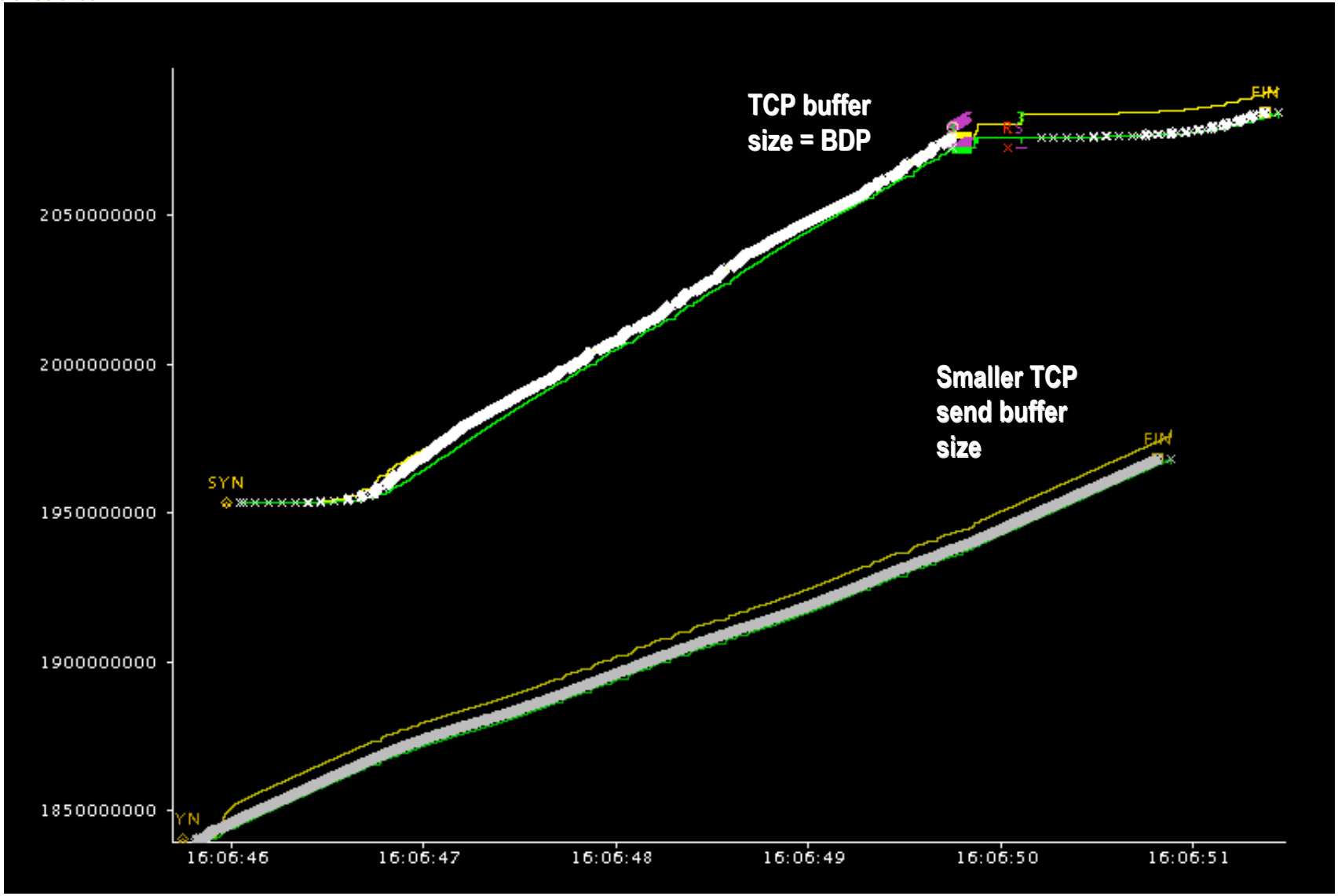
BERKELEY LAB



U.S. Department of Energy
Office of Science

Results: TCP throttling

BERKELEY LAB



Deployment

- Current installations
  - National Energy Research Scientific Computing Center (NERSC)
  - Lawrence Berkeley National Lab (LBNL)
  - Oak Ridge National Lab (ORNL)
  - Stanford Linear Accelerator Center (SLAC)

U.S. Department of Energy
Office of Science

Current / Future Work

BERKELEY LAB

- **Deployment at more sites**
- **Network administrator mode**
  - **Use signed and authorized activation packets to allow the ability to:**
    - Activate monitoring from a host that is not one of the endpoints
    - Send packet headers to a host that is not one of the endpoints
- **Capture at 10 GigE rate**

U.S. Department of Energy
Office of Science

Further Information

BERKELEY LAB

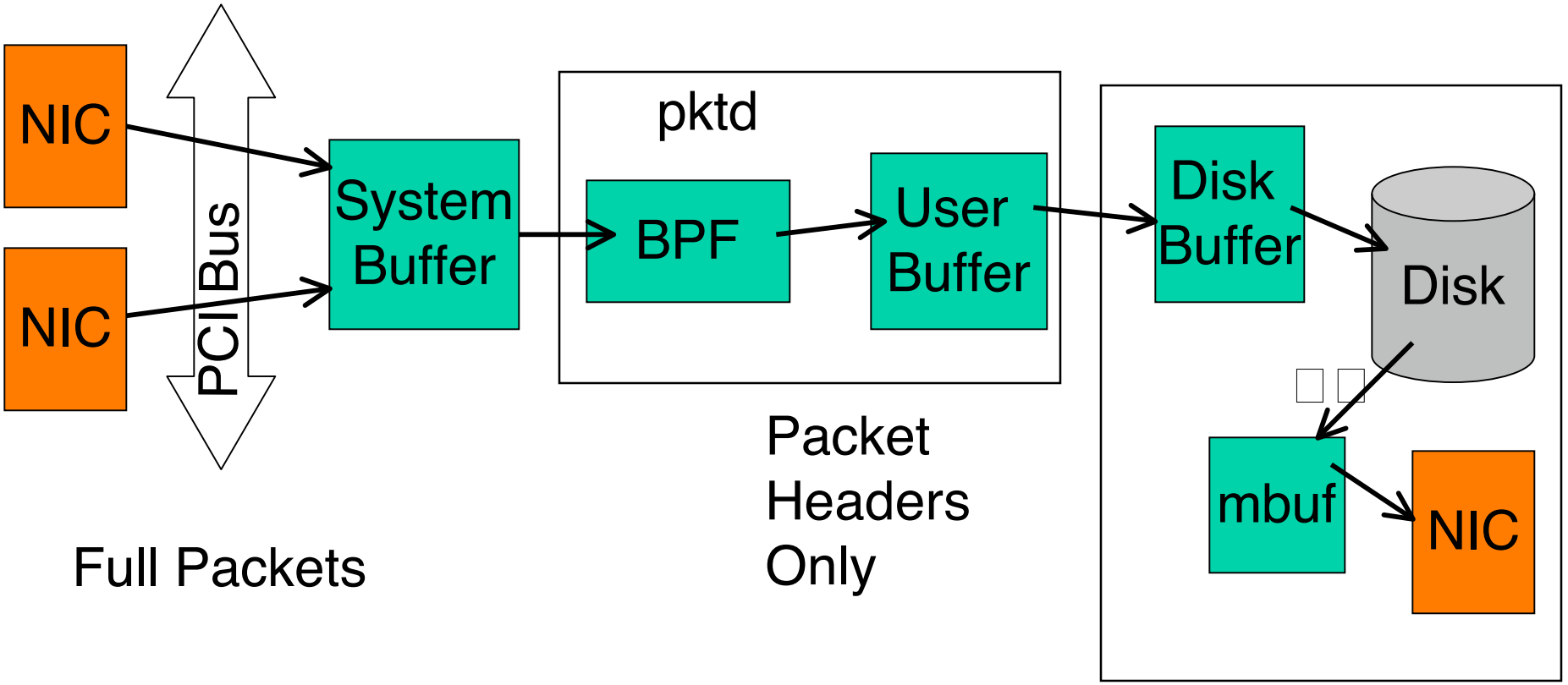
- URL
  - http://www-didc.lbl.gov/SCNM/
- Contacts
  - BLTierney@lbl.gov
  - DAAgarwal@lbl.gov
  - j_guojun@lbl.gov
  - chema@cs.berkeley.edu

Extra Slides

U.S. Department of Energy
Office of Science

Data Path

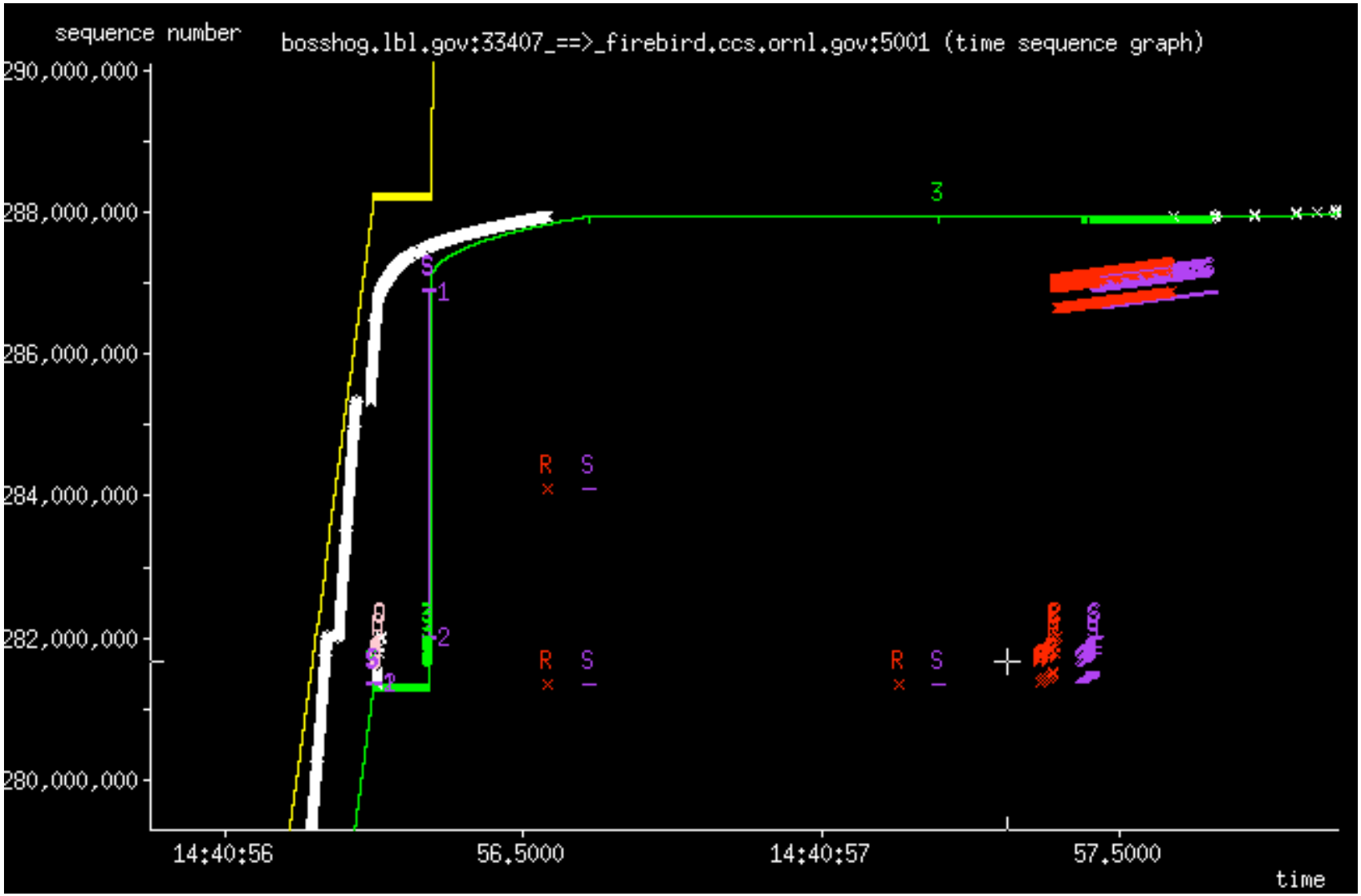
BERKELEY LAB



U.S. Department of Energy
Office of Science

Another view of the Linux SACK Bug

BERKELEY LAB



More on Interrupt Moderation

- **The maximum interrupt interval depends on**
  - **average packet size**
  - **space the kernel reserves for incoming data from the NIC**
    - number of per-packet kernel receive buffers (N)
  - **The maximum interrupt interval can thus be calculated as:**
    - time = N * average_packet_size / line_speed
  - **Defaults for the SysKonnect driver**
    - 256 kernel receive buffers
    - 200 ms interrupt servicing frequency
  - **For an average packet size of 1000 bytes:**
    - 256 receive buffers will become occupied after 2 ms.
  - **If the packet size is l500-bytes, the interrupt interval can be as large as 3 ms**
  - **Increasing the number of receive buffers will allow a longer coalescence period**
    - however, a large number of receive buffers requires large system memory resources